HITIQA: Scenario Based Question Answering

Sharon Small, Tomek Strzalkowski, Tracy Janack, Ting Liu,
Sean Ryan, Robert Salkin, Nobuyuki Shimizu
The State University of New York at Albany
1400 Washington Avenue
Albany, NY 12222
{small,tomek,tj5550,tl7612,seanryan,rs6021,ns3203}@albany.edu

Paul Kantor, Diane Kelly, Robert Rittman, Nina Wacholder
Rutgers University
New Brunswick, New Jersey 08903
{kantor, nina, diane, rritt}@scils.rutgers.edu

Boris Yamrom
Lehman College of the City University of New York
Bronx, New York 10468
byamrom@lehman.cuny.edy

Abstract

In this paper we describe some preliminary results of qualitative evaluation of the answering system HITIQA (High-Quality Interactive Question Answering) which has been developed over the last 2 years as an advanced research tool for information analysts. HITIQA is an interactive open-domain question answering technology designed to allow analysts to pose complex exploratory questions in natural language and obtain relevant information units to prepare their briefing reports in order to satisfy a given scenario. The system uses novel data-driven semantics to conduct a clarification dialogue with the user that explores the scope and the context of the desired answer space. The system has undergone extensive hands-on evaluations by a group of intelligence analysts representing various foreign intelligence services. This evaluation validated the overall approach in HITIQA but also exposed limitations of the current prototype.

1 Introduction

Our objective in HITIQA is to allow the user to submit exploratory, analytical questions, such as "What has been Russia's reaction to U.S. bombing of Kosovo?" The distinguishing property of such questions is that one cannot generally anticipate what might constitute the answer. While certain types of things may be expected (e.g., diplomatic statements), the answer is heavily conditioned by what information is in fact available on the topic, background knowledge of the user, context in the scenario, intended audience, etc. From a practical viewpoint, analytical questions are often underspecified, thus casting a broad net on a space of possible answers. Therefore, clarification dialogue is often needed to negotiate with the user the exact scope and intent of the question, and clarify whether similar topics found might also be of interest to the user in order to complete their scenario report. This paper will present results from a series of evaluations conducted in a series of workshops with the intended end users of HITIQA (professional intelligence analysts) using the system to solve realistic analytic problems.

HITIQA project is part of the ARDA AQUAINT program that aims to make significant advances in the state of the art of automated question answering. In this paper we focus on our approach to analytical question answering in order to produce a report in response to a given scenario. We also report on the user evaluations we conducted and their results with respect to our unique approach.

2 Analytical QA Scenarios

Analytical scenarios are information task directives assigned to analysts to support a larger foreign policy process. Scenarios thus contain the information need specifications at various levels of detail, the type, format and timing of the response required (an intelligence report) as well as the primary recipient of the report (e.g., the Secretary of State). A hypothetical, but realistic scenario is shown in Figure 1 below. This scenario, along with several others like it, was used in evaluating

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2004		**00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
HITIQA: Scenario Based Question Answering	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **The State University of New York at Albany,1400 Washington Avenue,Albany,NY,12222**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**8**	

HITIQA performance and fitness for supporting the analytical process.

As can be readily assessed from the directives in Figure 1, scenarios are not merely tough questions; they are far too complex to be considered as a single question at all. It is equally clear that no simple answer can be expected and that preparing a report would mean finding answers to a series of interlocking questions or various granularities.

Scenario: The al-Qaida Terrorist Group

As an employee of the Central Intelligence Agency, your profession entails knowledge of the al-Qaida terrorist group. Your division chief has ordered a detailed report on the al-Qaida Terrorist Group due in three weeks. Provide as much information as possible on this militant organization. Eventually, this report should present information regarding the most essential concerns, including who are the key figures involved with al-Qaida along with other organizations, countries, and members that are affiliated, any trades that al-Qaida has made with organizations or countries, what facilities they possess, where they receive their financial support, what capabilities they have (CBW program, other weapons, etc.) and how have they acquired them, what is their possible future activity, how their training program operates, who their new members are. Also, include any other relevant information to your report as you see fit.

FIGURE 1: Scenario used during user evaluations

We have organized a series of usability evaluations with active duty intelligence analysts to find out how they approach the problem of solving a scenario. The prerequisites for this were are follows:

1. A robust, broadly functional analytical QA system capable of sustaining realistic analytic tasks.
2. A realistic corpus of "raw intelligence" in form of varying quality and verity new-like reports.
3. A set of realistic, average complexity analytic tasks or scenarios to be used.

HITIQA has been developed over the past two years as an open-ended highly flexible interactive QA system to allow just this type of evaluation. The system supports a variety of information gathering functions without straight jacketing the user into any particular mode or interaction style. The system does not produce cut and dry "answers"; instead it allows the analysts to build the answers the way they want them. While this open-endedness may seem like unfinished business, we believe that further development must take into account the needs of analysts if they were ever to adopt this technology in their work.

Our main hypothesis is that analysts employ a range of strategies to find the required information and that these strategies depend significantly upon the nature of the task and the progress the analyst is making on the task, in addition to individual differences between analysts. Our experience with interactive systems also indicated that real users are unlikely to follow any single information exploration strategy, but instead would use multiple, parallel, even overlapping approaches in order to maximize the returns and their confidence in the results. As a corollary we may expect that the scenario tasks are unlikely to be systematically decomposed into a series of smaller tasks *ahead* of actual search. In other words, the analytical process is a dialogue, not a sequence of commands. Moreover, questions actually submitted to the system during the analytical process seldom seek just the exact answer, instead they are often considered as "light beams" through the data: focusing on the answer but also illuminating adjacent, related information which may prove just as valuable.

AFRL, NIST, CNS and ARDA collaborated in the development of scenarios used in our evaluation sessions.

3 Data Driven Semantics of Questions

When the user poses a question to a system having access to a huge database of unstructured data (text files), we need to first reduce the big pile to perhaps a handful of documents where the answer is likely to be found. The easiest way to do it is to convert the question into a search query (by removing stopwords and stemming and tokenizing other words) and submitting this query to a fast but non-exact document retrieval system, e.g., Smart (Buckley, 1985) or InQuery (Callan et al., 1992), or if you are on the web, Google, etc.

In the current prototype of HITIQA, we use a combination of Google and InQuery to retrieve the top 50 to 200 documents from a large document database, consisting of several smaller collections such as newspaper stories, documents from the Center of Nonproliferation Studies, as well as web mined files. The retrieved documents are then broken down into passages, mostly exploiting the naturally occurring paragraph structure of the original sources.

The set of text passages returned from the initial search is the first (very crude) approximation of the *Answer Space* for the user's first question. In order to determine what this answer space consists of we perform automatic analysis (a combination of hierarchical clustering and classification) to uncover if what we got is a fairly homogenous collection (i.e., all texts have very similar content), or whether there are a number of diverse topics or aspects represented in there, somehow tied together by a common thread. In the former case, we may be reasonably confident that we have the answer, modulo the retrievable information. In the latter case, we know that the question is more complex than the user may have intended, and a negotiation process is needed to clarify topics of interest for the scenario report.

The next step is to measure how well each of the aspects within the answer space is "matching up" against the original question. This is accomplished through the framing process described later in this paper. The outcome of the framing process is twofold: first, the alternative interpretations of the question are ranked within 3 broad categories: on-target, near-misses and outliers. Second, salient concepts and attributes for each topical/aspectual group are extracted into topic frames. This enables the system to conduct a meaningful dialogue with the user, a dialogue which is wholly content oriented, and entirely data driven.

4 Partial structuring of text data

In HITIQA we use a text framing technique to delineate the gap between the meaning of the user's question and the system "understanding" of this question. The framing is an attempt to impose a partial structure on the text that would allow the system to systematically compare different text pieces against each other and against the question, and also to communicate with the user about this. In particular, the framing process may uncover topics or aspects within the answer space which the user has not explicitly asked for, and thus may be unaware of their existence. This approach is particularly beneficial to the needs of the scenario problem, where these similar aspects frequently are needed in completely "answering" the scenario, with the scenario report.

In the current version of HITIQA, frames are predefined structures representing various *event types*. We started with the General frame, which can represent any event or relation involving any number of entities such as people, locations, organizations, time, and so forth. In a specialized domain, or if the user interests are known to be limited to a particular set of topics, we define domain-specific frames. Current HITIQA prototype has three broad domain-specific frames, related to the Weapon of Mass Destruction proliferation domain (which was one of the domains of interest to our users). These frames are: *WMDTransfer, WMDDevelop*, *WMDTreaty,* and of course we keep the *General* frame. Obviously, these three frames do not cover the domain represented by our data set; they merely capture the most commonly occurring types of events. All frames contain a small number of core attributes, such as LOCATION, PERSON, COUNTRY, ORGANIZATION, ETC., which are extracted using BBN's Identifinder software, which extracts 24 types of entities. Domain-specific frames add event specific attributes, which may require extracting additional items from text, or assigning *roles* to existing attributes, or both. For example, WMDTransfer's attributes TRANSFER_TO and TRANSFER_FROM define roles of some COUNTRY or ORGANIZATION, while the TRANSFER_TYPE attribute scans the text for keywords that may indicate the type of transfer, e.g., *export, sale*, etc.

HITIQA creates a *Goal* frame for the user's question, which can be subsequently compared to the data frames obtained from retrieved data. A Goal frame can be a General frame or any of the domain specific frames available in HITIQA. For example, the Goal frame generated from the question, *"Where does al-Qaida have training facilities*?" is a General frame as shown in Figure 2. This was the first question generated by one of our analysts during the first evaluation while working on the al-Qaida scenario shown in Figure 1.

FRAME TYPE: *General*
TOPIC: *training facilities*
ORGANIZATION: *al-Qaida*

FIGURE 2: HITIQA generated General-type Goal frame from the al-Qaida training facilities question

FRAME TYPE: *General*
CONFLICT SCORE*: 1*
TRANSFER TYPE: *provided*
TRANSFER TO: *al-Qaida*
TRANSFER FROM: *Iraq*
TOPIC: *provided*
SUB-TOPIC: *imported*
LOCATION: *Iraq*
PEOPLE: *Abu Musab al-Zarqawi, Bush, George Tenet, Saddam Hussein*
ORGANIZATION:*CIA, Administration, al-Qaida*
DOCUMENT: *web_283330*
PARAGRAPHS: ["***CIA*** *chief* ***George Tenet*** *seems to have gone a long way to back the* ***Bush Administrations*** *declarations that the long split between Islamic fundamentalist terrorist organizations like* ***Al-Qaida*** *and secular Iraqi ruler* ***Saddam Hussein*** *is healed.*

He has testified that the ***CIA*** *has evidence of* ***Iraqi*** *providing* ***Al Qaida*** *with training in forgery and bomb making and of providing two,* ***Al Qaida*** *associates with training in gas and poisons. He said also that* ***Iraq*** *is harboring senior members of a terrorist network led by* ***Abu Musab al-Zarqawi****, a close* ***Al Qaida*** *associate.* "]
RELEVANCE: *Conflict: [Topic]*

FIGURE 3: A HITIQA generated data frame and the underlying text passage. Words in bold were used to fill the Frame.

HTIQA automatically judges a particular data frame as relevant, and subsequently the corresponding segment of text as relevant, by comparison to the Goal frame. The data frames are scored based on the number of conflicts found between them and the Goal frame. The conflicts are mismatches on values of corresponding attributes. If a data frame is found to have no conflicts, it is given the highest relevance rank, and a conflict score of zero. All other data frames are scored with

a decreasing value based on the number of conflicts, negative one for frames with one conflict with the Goal frame, negative two for two conflicts etc. Frames that conflict with all information found in the question are given a score of -99 indicating the lowest relevancy rank. Currently, frames with a conflict score of -99 are excluded from further processing as outliers. The frame in Figure 2 is scored as a near miss and will generate dialogue, where the user will decide whether or not it should be included in the answer space.

5 Clarification Dialogue

Data frames with a conflict score of zero form the initial kernel answer space. Depending upon the presence of other frames outside of this set, the system either proceeds to generate the answer or initiates a dialogue with the user. HITIQA begins asking the user questions on these near-miss frame groups, with the largest group first. The groups must be at least groups of size N, where N is a user controlled setting. This setting restricts all of HITIQA's generated dialogue.

A one conflict frame has only a single attribute mismatch with the Goal frame. This could be a mismatch on any of the General attributes, for example, LOCATION, or ORGANIZATION, or TIME, etc., or in one of the domain specific attributes, TRANSFER_TO, or TRANSFER_TYPE, etc. A special case arises when the conflict occurs on the TOPIC attribute. Since all other attributes match, we may be looking at potentially different events or situations involving the same entities, or occurring at the same location or time. The purpose of the clarification dialogue in this case is to probe which of these topics may be of interest to the user. Another special case arises when the Goal frame is of a different type than a data frame. The purpose of the clarification dialogue in this case is to expand the user's answer space into a different but possibly related event. A combination of both of these cases is illustrated in the exchange in Figure 4 below.

User: *"Where does al-Qaida have training facilities?"*

HITIQA: *"Do you want to see material on the transfer of weapons and intelligence to al-Qaida?"*

FIGURE 4: Dialogue generated by HITIQA for the al-Qaida training facilities question

In order to understand what happened here, we need to note first that the Goal frame for this example is a General Frame, from Figure 2. One of the data frames that caused this dialogue to be generated is shown in Figure 3 above. While this frame is of a different frame type than the Goal frame, namely WMD Transfer, it matches on all of the General attributes except TOPIC, so HITIQA asks the user if they would like to expand their answer space to this other domain, namely to include the transfer of weapons involving this organization as well.

ANSWER REPORT:

The New York Times said the Mindanao had become the training center for the Jemaah Islamiah network, believed by many Western governments to be affiliated to the **al-Qaida** movement of Osama bin Laden
DocName: A-web_283305 ParaId: 2

…
IRAQ REPORTED TO HAVE PROVIDED MATERIALS TO AL QAIDA
2003
[CIA chief George Tenet seems to have gone a long way to back the Bush Administrations declarations that the long split between Islamic fundamentalist terrorist organizations like Al Qiada and secular Iraqi ruler Saddam Hussein is healed.
DocName: A-web_283330 ParaId: 6
He has testified that the CIA has evidence of Iraqi providing Al Qaida with training in forgery and bomb making and of providing two, Al Qaida associates with training in gas and poisons. He said also that Iraq is harboring senior members of a terrorist network led by Abu Musab al-Zarqawi, a close Al Qaida associate. The Bush Administration and the press has carelessly shorthanded this to mean, a senior Al Qaida member, ignoring the real ambiguities that surround the true nature of that association, and whether Zarqawi shares Al Qaidas ends, or is receiving anything more than lodging inside Iraq.]
DocName: A-web_283330 ParaId: 7

FIGURE 5: Partial answer generated by HITIQA to the al-Qaida training facilities question

During the dialogue, as new information is obtained from the user, the Goal frame is updated and the scores of all the data frames are reevaluated. The system may interpret the new information as a positive or negative. Positives are added to the Goal frame. Negatives are stored in a Negative-Goal frame and will also be used in the re-scoring of the data frames, possibly causing conflict scores to increase. If the user responds the equivalent of "yes" to the system clarification question in Figure 4, a corresponding WMD Transfer frame would be added to the Goal frame and all WMD Transfer frames will be re-scored. If the user responds "no", the Negative-Goal frame will be generated and all WMD Transfer frames will be rescored to 99 in order to remove them from further processing. The user may end the dialogue, at any point and have an answer generated given the current state of the frames.

Currently, the answer is simply composed of text passages from the zero conflict frames. In addition, HITIQA will generate a "headline" for the text passages in all the Frames in the answer space. This is done using grammar rules and the attributes of a frame. Figure

5 shows a portion of the answer generated by HITIQA for the al-Qaida training facilities question.

6 HITIQA Interface

There are two distinct ways for the user to interact with HITIQA to explore their answer space. The Answer Panel displays the user's current answer at any given time during the interaction for a single question. Through this panel the user can read the paragraphs that are currently in their answer. There are links on this panel so the user is able to view the full original source document from which the passage(s) were extracted.

The Visual panel offers the user an alternative to reading text by providing a tool for visually browsing the entire answer space. Figure 6 shows a typical view of the visualization panel. The spheres are representative of single frames and groups of frames. The user's attention may be drawn to particular frames by the color coding or the attribute spikes. The colors represent the frame's score, so the user can quickly see what is in their answer, blue, and what is not, all other colors. The attribute spikes may also be used as a navigation tool. The active attribute is chosen by the user through radio buttons. The current active attribute in Figure 6, is Location. This displays all instances of locations mentioned in the corresponding text.

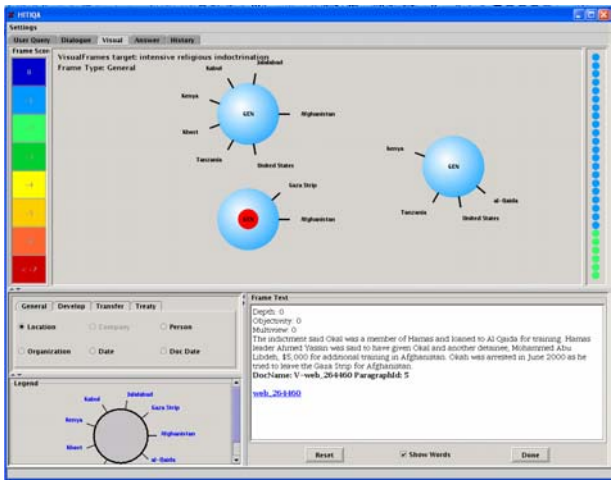

Figure 6: Frame Level Display

The underlying text that was used to build the frame may be displayed in the lower right hand window. In this text display window there is a hyperlink that takes the user directly to the full source document. The user is able to interact with this panel by adding and removing information from their generated answer. Moving from the visualization to the textual dialogue, the generated answer, and back is seamless in a sense that any changes to the frame scores in one modality are immediately accessible to the user in another modality. Users can add and remove frames from the answer space and HITIQA will always seamlessly pickup a new dialogue or generate a new answer.

7 HITIQA Qualitative Evaluations

In order to assess our progress thus far, and to also develop metrics to guide future evaluation, we invited a group of analysts employed by the US government to participate in two three-day workshops held in September and October 2003.

The two basic objectives of the workshops were:

1. To perform a realistic assessment of the usefulness and usability of HITIQA as an end-to-end system, from the information seeker's initial questions to completion of a draft report.

2. To develop metrics to compare the answers obtained by different analysts and evaluate the quality of the support that HITIQA provides.

Each of these objectives entails a particular challenge. Performing a realistic assessment of HITIQA is difficult because many of the resources that the analysts use, as well as the reports they produce, are classified and therefore inaccessible to researchers.

Assessing the quality of the support that the system provides is not easy because analytical questions rarely have a single right answer. It is not obvious how to define, for example, the precision of the system. We therefore conducted an 'information unit' exercise, whose purpose was to determine whether the analysts could identify information building blocks in their reports, so that we could compare and contrast different reports.

To obtain an adequate supply of appropriate text data to support extensive question answering sessions (1, 2, 3 and 4 hours long), we prepared a new corpus of approximately 1.2 Gbytes. This new corpus consists of the reports from the Center for Non-Proliferation Studies (CNS) collected for the AQUAINT Program, augmented with a much larger collection of texts on similar subject matter mined from the web using Google[1]. The final corpus proved to be sufficient to support about three hours of use of HITIQA to "solve" each of the scenarios.

The first day of the first workshop was devoted to training, including a two-part proficiency test. HITIQA is a fairly complex system, that includes multiple layers of data processing and user interaction, and it was critical that the users are sufficiently "fluent" if we were to measure their productivity. The analysts' primary task on the second day was preparation of reports in response to the scenarios.

[1] Google has kindly agreed to temporarily extend our usage license so we could collect the data over a short time.

The third day was devoted to quantitative and qualitative evaluation, discussed later. In addition, we asked the analysts to score each others reports, as well as to identify key information units in them. These information units could be later compared across different reports in order to determine their completeness.

8 Workshop Results

The results of the quantitative evaluations strongly validate the approach that we have taken. These conclusions are confirmed by analysts comments gleaned both from the formal qualitative assessment and from informal discussion. As one analyst said, "the system as it stands now, in my mind, gave me enough information to try to put together a 80% solution but …I don't think you're ever gonna reach that 100% state." At the same time, we learned a great deal about how analysts work.

It is important to determine the realism of the scenarios used during the workshop relative to the analysts' current work tasks in order for any results to be meaningful. Each analyst was asked a series of five questions such as, "*How realistic was the scenario? In other words did it resemble tasks you could imagine performing at work?*" These 5 questions were all relative to the realism and difficulty of the scenario tasks. Analysts used a scale of 1 to 5 based on their agreement with the statements, where 5 was complete agreement. Our mean score was 3.84, indicating our scenarios were realistic and of about average difficulty when compared to the work they normally perform.

We have classified the type of passages that an analyst copied to their report into two categories, answer passages and additional information passages, see Figure 7 below. The answer passages either exactly answered the user's initial question or supplied supporting information. The additional passages do not answer the original question posed, but may have been added to the answer through dialogue, or through the user's exploration of document links offered. This could be a piece of information needed to satisfy some other aspect of the scenario that they had not asked about yet, or possibly a topic the user had not even considered but found relevant when it was presented to them. As can be seen there was a very large amount of "additional" information that the user copied to their report. The amounts reported here are the averages for all of the analysts for both workshops. This supports our hypothesis that analysts seldom seek just the exact answer, but they are also looking at adjacent, related information, much of which they retain for their report. Note that there were a small number of passages that contained a combination of answer and additional information; these were added to answer.

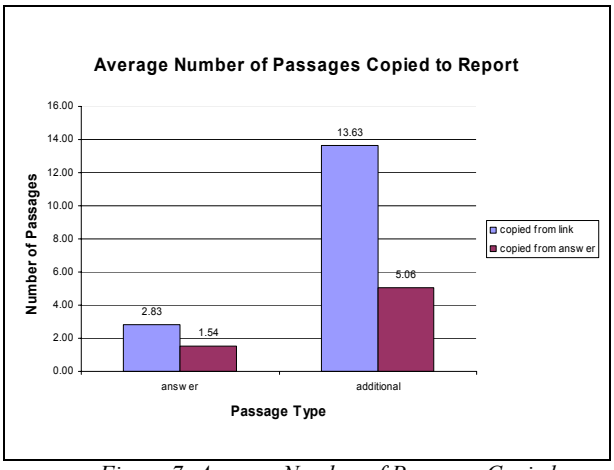

Figure 7: Average Number of Passages Copied

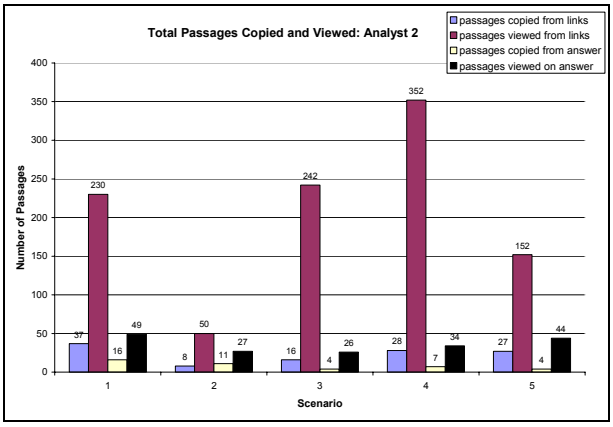

Figure 8: Number of Passages Copied Vs. Those Viewed

We should now establish the number of passages copied versus those viewed, relative to links and the answer. Figure 8 above shows the total number of passages copied versus the total number of passages viewed. It is seen that many more passages need to be viewed through full document links before a useful passage is found. In comparison a much smaller number of answer passages need to be viewed from the Answer panel in order to find useful passages.

All of the analysts' sessions were recorded using Camtasia. Figure 9 shows an annotation created for a typical session. Analysts were observed to utilize a range of varying strategies as they worked different scenarios and even while working different queries of the same scenario. Figure 10 shows the statistics for each Analyst's use of HITIQA while working on the scenarios during the two workshops (note that Analyst-4 was only able to attend the first workshop and Analyst-1 did not create a report for Scenario 2). Some of the variations in strategies among the analysts while working the same scenario are quite striking. For example, Scenario 4 was worked quite differently by Analyst-1

versus Analyst-2. While Analyst-1 spent almost all of his/her time in the Visual Panel, Analyst-2 spent virtually all of his/her time in the Answer panel. Analyst-1 produced his/her report copying 52 paragraphs while Analyst 2 copied only 35. There are also large variations in the number of questions asked for the same scenario. Examine scenario 5, where Analyst-3 asked a total of 11 questions and Analyst-2 only needed to ask 2 questions. Relative to this, Analyst-3, who asked a much larger number of questions, copied only 28 passages, whereas Analyst-2 copied 31. These variations, as stated earlier in the paper, could be due to the nature of the task, the progress the analyst is making on the task, in addition to individual differences between analysts. For example, the difference in the number of questions asked between Analyst-2 and Analyst-3 for scenario 5 may be due to difference in search strategies employed, but may also reflect the amount of background knowledge of the topic.

User: *What is the status of South Africa's chemical, biological, and nuclear programs?*
Clarification Dialogue: 1 minute
- 6 questions generated by HITIQA
  - replied "Yes" to 5 and "No" to 1
  - 5+ passages added to answer

Studying Answer Panel: 60 minutes
- Copying 24 passages to report
  - 10 from Answer
  - 14 from Links to Full Document
- Visual Panel Browsing: 5 minutes
  - Nothing copied

User: *Has South Africa provided CBW material or assistance to any other countries?*
Clarification Dialogue: 1 minute
- 5 questions generated by HITIQA
  - replied "Yes" to 2 and "No" to 3
  - 2+ passages added to answer

Studying Answer Panel: 26 minutes
- Copying 6 passages to report
  - 6 from Links to Full Document

Visual Panel browsing: 1 minute
- Copying 1 passage to report
  - 1 from Links to Full Document

User: *How was South Africa's CBW program financed?*
Clarification Dialogue: 40 seconds
- 7 questions generated by HITIQA
  - replied "Yes" to 3 and "No" to 4
  - 3+ passages added to answer

Studying Answer Panel: 11 minutes
- Copying 3 passages to report
  - 1 from Answer

2 from Links to full Document

FIGURE 9: Fragment of an analytical session

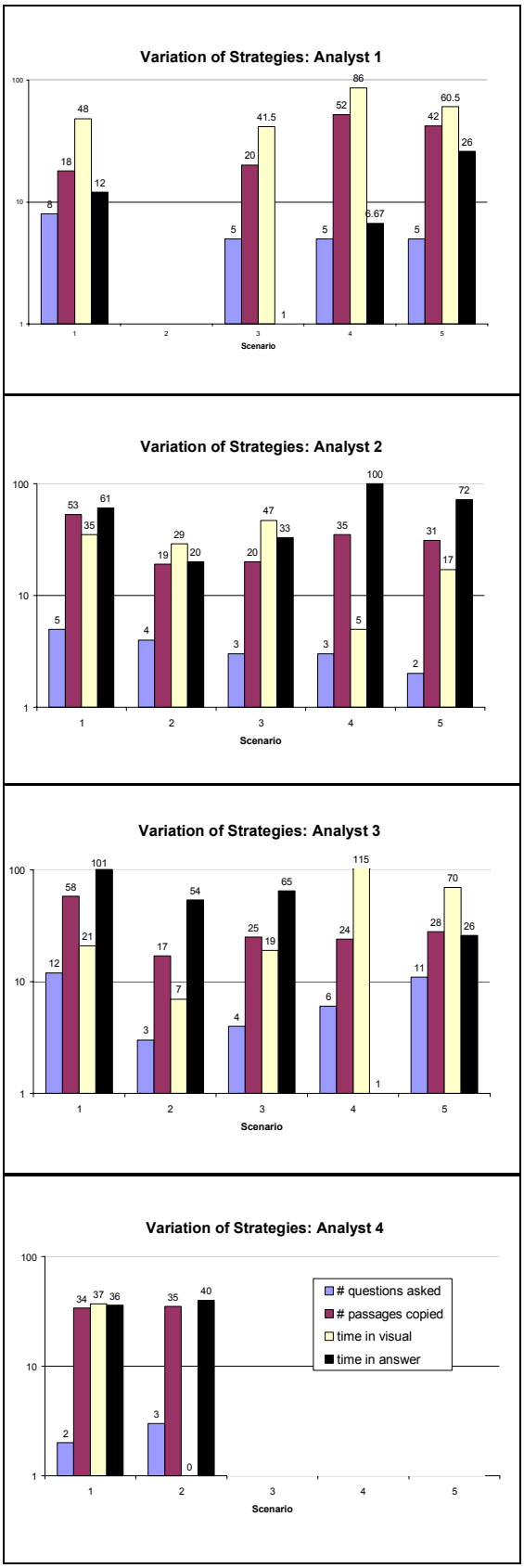

Figure 10: Varying Strategies Employed

There is, however, some consistency across the analysts in the amount of information retained per scenario. The charts are drawn in logarithmic scale, but it should be visible that scenarios 2 and 3 produced less interaction and required less information to fulfill than scenarios 4 and 5. It is also visible that scenario 1 required more questions to be asked and more exploration to be done in visual panel than other scenarios.

Finally, it is important to provide some metric regarding the user's overall satisfaction with their use of HITIQA. At the end of each workshop Analysts were given a series of 17 questions, such as "*HITIQA helps me find important information*", shown in Figure 11, to assess their overall experience with the system. Many of these questions were designed for the user to compare HITIQA to the current tools they are using for this type of task. Analysts again used a scale of 1 to 5 based on their agreement with the statements. The results were then converted, where 5 would always denote the best, and are shown in Figure 11 below. It is important to note that we scored highly overall, but additionally we scored highly in the majority of questions relative to comparison of their current tools. For example, for Question 14: *"Having HITIQA at work would help me find information faster than I can currently find it",* our mean score was 3.83.

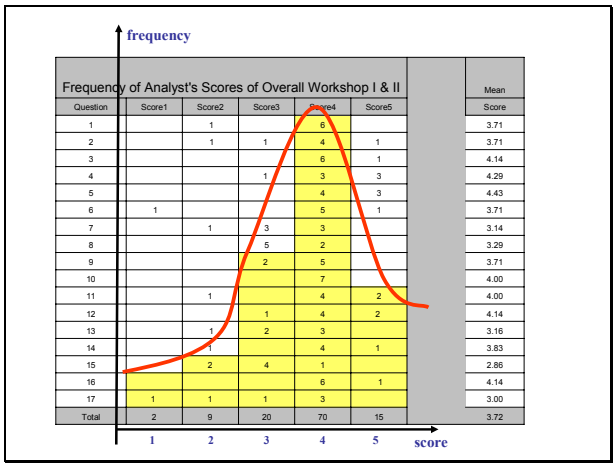

FIGURE 11: Final Evaluation Results, Workshop 1 & 2

In summary, the results from these two evaluations indicate that HITIQA, in its current state, is already competitive with the tools that the analysts are currently using in their work, supporting our overall approach to Analytical Question Answering. HTIQA provides the user with a tool to find the passages needed to complete a report for a given scenario. While working on a scenario HITIQA has been shown to provide information which exactly answers the user's question, and additionally HITIQA's method brings to light other related information that the analyst retains in order to complete their report.

Acknowledgements

This paper is based on work supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program under contract number 2002-H790400-000.

References

Allen, J. and M. Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. www.cs.rochester.edu/research/cisd/

Baeza-Yates and Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley.

Chris Buckley. 1985. *Implementation of the Smart information retrieval system.* Technical Report TR85-686, Department of Computer Science, Cornell University, Ithaca, NY.

Ferguson, George and James Allen. 1998. *TRIPS: An Intelligent Integrated Problem-Solving Assistant*, in Proceedings of the 15th AAAI Conference (AAAI-98), Madison, WI, pp. 567-573.

Hardy, H., N. Shimizu, T. Strzalkowski, L. Ting, B. Wise and X. Zhang. 2002a. *Cross-Document Summarization by Concept Classification*. Proceedings of SIGIR, Tampere, Finland.

Hardy, H., K. Baker, L. Devillers, L. Lamel, S. Rosset, T. Strzalkowski, C. Ursu and N. Webb. 2002b. *Multi-layer Dialogue Annotation for Automated Multilingual Customer Service*. ISLE Workshop, Edinburgh, Scotland.

Harabagiu, S., et. al. 2002. *Answering Complex, List and Context questions with LCC's Question Answering Server.* In Proceedings of Text Retrieval Conference (TREC-10).

Hovy, E., L. Gerber, U. Hermjakob, M. Junk, C-Y. Lin. 2000. *Question Answering in Webclopedia. Notebook*. Proceedings of Text Retrieval Conference (TREC-9).

Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, Y. Wilks. 1998. Description of the LaSIE-II System as Used for MUC-7. In Proceedings of the Seventh Message Understanding Conference (MUC-7.)

Litman, Diane J. and Shimei Pan. 2002. *Designing and Evaluating an Adaptive Spoken Dialogue System*. User Modeling and User-Adapted Interaction. Vol. 12, No. 2/3, pp. 111-137.

Seneff, S. and J. Polifroni. 2000. *Dialogue Management in the MERCURY Flight Reservation System*. Proc. ANLP-NAACL 2000, Satellite Workshop, pp. 1-6, Seattle, WA.

Small, Sharon, Nobuyuki Shimizu, Tomek Strzalkowski and Liu Ting (2003). *HITIQA: A Data Driven Approach to Interactive Question Answering: A Preliminary Report*. AAAI Spring Symposium on New Directions in Question Answering, Stanford University, March 24-26, 2003. pp. 94—104.

Tang, Rong, K.B. Ng, Tomek Strzalkowski and Paul Kantor (2003). Automatic Prediction of Information Quality in News Documents. Proceedings of HLT-NAACL 2003, Edmonton, May 27-June 1

Walker, Marilyn A. 2002. *An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email* . Journal of AI Research, vol 12., pp. 387-416.